Redesign des AddressCheckers der Post CH AG

Fachgebiet: Enterprise Application Development / Full-Text Search
Betreuer: Bruno Reist (Post CH AG)
Experte: Stephan Fischli (BFH)
Industriepartner: Post CH AG, Bern

Um diverse Dienstleistungen und Produkte im Adressumfeld anzubieten, unterhält die Post CH AG im Rahmen ihres Service Public Auftrages eine der aktuellsten und vollständigsten Adressdatenbanken der Schweiz. Den sehr hohen Anforderungen an die Suche wird mit einer clusterbasierten Volltextsuch-Engine Rechnung getragen. Um die Kundenbedürfnisse noch besser abzudecken, wird eine neue, mehrstufige Adressvalidierungslogik auf Basis von Elasticsearch konzipiert und implementiert.

Umfeld

Die Post CH AG betreibt ein zentrales Adressmanagement-System, welches unter anderem die Datengrundlage für die Brief- und Paket-Sortierung liefert. AMPplus ist ein strategisch wichtiges System, welches als primäre Anlaufstelle für Adressauskünfte anderer Systeme dient und somit hohe Anforderungen an die Schnittstellen zu erfüllen hat. Über den AddressChecker-Webservice wird eine Validierung von Eingabedaten gegenüber dem gesamten Adressdatenbestand der Post in Echtzeit ermöglicht.

Ausgangslage

Der AddressChecker-Webservice liefert Endbenutzern Informationen, ob die Zustellung an eine bestimmte Adresse möglich ist oder nicht. Die bestehende Lösung basiert dabei auf einem einfachen Verfahren, welches die Benutzereingaben lediglich mit exakten Textalgorithmen abgleicht und anschliessend eine Rückmeldung liefert. Die Tatsache, dass diese Algorithmen bereits bei kleinsten Abweichungen zwischen Eingabe- und Referenzdokument keine Treffer mehr liefern, entspricht nicht den heutigen Erwartungen an ein Validierungs- oder Suchsystem. Damit der Validierungsservice trotz Tippfehlern oder bei ähnlichen Schreibweisen dennoch die korrekten Resultate liefert, müssen zusätzlich verschiedene computerlinguistische Algorithmen in möglichst effizienter Art und Weise angewendet werden. Eine komplett neue, umfangreichere Validierungslogik soll Abhilfe für die obengenannte Problematik schaffen und die Adressvalidierung für Kunden der Post CH AG auf einen zeitgemässen Stand bringen.

Lösung

Neben den Vorteilen, welche sich aus dem Einsatz von unscharfen Textalgorithmen ergeben, bergen diese auch die Gefahr, dass Sucheingaben ihre Eindeutigkeit verlieren. Diese Umstände verlangen ein stufenweises Vorgehen. In einer ersten Stufe wird die Eingabe zuerst exakt abgeglichen. Ergibt das Ergebnis eine leere Resultatmenge, wird die Suche mit unscharfen Methoden fortgesetzt. Eine weitere Hürde stellt der umfassende Lifecycle einer Adresse dar. Dieser führt dazu, dass ein Minimum von neun Volltextabfragen notwendig werden, um sämtliche Zustände der Validierungsantwort abzubilden. Damit die Effizienz erhöht und die Antwortzeit generell gesenkt werden kann, werden die Abfragen anhand ihrer Laufzeit bzw. Komplexität gruppiert und in Phasen parallel ausgeführt. Die Validierung von korrekten und vollständigen Eingaben kann so bereits nach der ersten Phase terminiert werden. Für die Ausführung und Steuerung dieser stufenweisen Validierung wurde eine phasenbasierte «Finite-state machine» entwickelt, welche einerseits die phasenweise Validierung erlaubt und andererseits die parallele Ausführung der notwendigen Volltextsuchanfragen unterstützt. Die durch hunderte Volltextsuchabfragen pro Sekunde über Millionen von Datensätzen entstehende Last erfordert eine beträchtliche Rechenleistung. Aus diesem Grund werden Suchlogik sowie Suchdaten auf eine Vielzahl von Rechnern aufgeteilt, so dass Antwortzeiten unter 100ms garantiert werden können. Durch geschicktes Sharding, einer Art Partitionierung, Verteilung der Suchindizes, sowie der Parallelisierung der Abfragen wird das Optimum aus Antwortzeit und Ausfallsicherheit unter Berücksichtigung der eingesetzten Hardware erzielt.

Fazit

Die Einführung der phasenbasierten «Finite-state machine» bringt enorme Vorteile mit sich, welche sich vor allem in erhöhter Performanz sowie verbesserter Wartbarkeit manifestieren. Die Implementation von drei Validierungsphasen, «aktiv», «aktiv-unscharf» und «historisch», ermöglicht eine extrem schnelle Beantwortung korrekter und vollständiger Validierungsanfragen. Unvollständige und/oder veraltete Adressen durchlaufen hingegen mehrere Phasen, welche im Endeffekt aber dennoch zu einem Validierungsresultat im Zehntelsekunden-Bereich führen. Die Kapselung zwischen der phasenbasierten «Finite-state machine» und den Validierungsschritten erlaubt ein unkompliziertes und gezieltes Eingreifen in die Businesslogik, welche nach modernen Methoden der objektorientierten Programmierung umgesetzt wurde.

Lukas Feuz
lukas.feuz@post.ch